

FIT3164 Data Science

Software Project 2

User Guide

Team	DS-02
Project title	Patient medical history summarisation
Team members	Rui Guo (30481872) Govind Chadha (31113567) Hamza Umar (31181465)

1. End User Guide	3
1.1 User Interface	3
a. Patient summaries	3
b. Save as a csv file	5
1.2. Limitations	7
a. Privacy of user	7
b. Update of data	7
c. The name of saved files	7
d. Accessibility	7
2. Technical Guide	8
2.1 Setup and Configuration	8
2.2. User Interface	8
a. Front-end code and relevant data	8
b. Run and Test	9
c. Limitation and Further Improvement	9
2.3 LLM (Large Language Model)	9
a. Model Details	9
b. Preprocessing Steps	10
c. Model Application	11
d. Output	12
e. Challenges and Solutions	12
2.4 Database	13
3. Acknowledgement	14

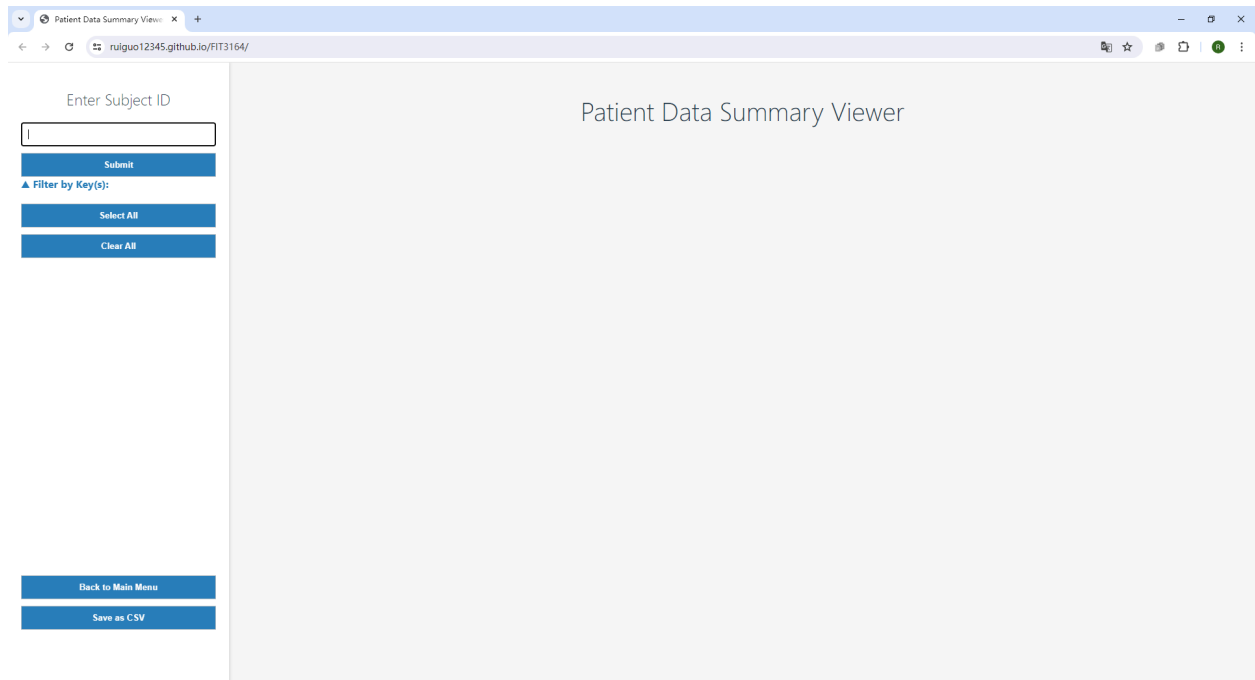
1. End User Guide

This document is a user guide on how to use the patient summary application. The final product contains two main features. This guide will explore in depth how to make the most of these two components, application limitations, and troubleshooting.

1.1 User Interface

a. Patient summaries

The first component allows the user to enter a subject id (patient id), click "submit," and the site will get a csv file (pre-processed using the LLM model) to view a summary of the patient records for this patient. Therefore, in the first step, confirm the patient's subject id and enter it into the "enter subject id" field. You will see a submit button below it. Please click this button and all patient data will be displayed in the blank space on the right side of the screen.



The screenshot shows a web browser window with the title "Patient Data Summary Viewer". The address bar shows the URL "rulguo12345.github.io/FIT3164/". The main content area is divided into two sections. On the left, there is a form with the label "Enter Subject ID" above a text input field. Below the input field is a blue "Submit" button. Underneath the submit button is a section labeled "▲ Filter by Key(s):" with two blue buttons: "Select All" and "Clear All". At the bottom of the left sidebar, there are two more blue buttons: "Back to Main Menu" and "Save as CSV". The right section of the main content area is a large, empty light gray box with the text "Patient Data Summary Viewer" centered at the top.

Figure 1, main menu

The screenshot shows a web application titled "Patient Data Summary Viewer". On the left, there is a sidebar with the following elements:

- Input field: "Enter Subject ID" with the value "10000032".
- Buttons: "Submit", "Filter by Key(s):", "Select All", and "Clear All".
- Buttons at the bottom: "Back to Main Menu" and "Save as CSV".

The main content area displays a table with the following data:

Key	Value
subject_id	10000032
Sex	F
Service	MEDICINE
Allergies	Percocet / Vicodin
Attending	REDACTED REDACTED. REDACTED REDACTED.
Chief Complaint	Worsening ABD distention and pain abdominal fullness and discomfort altered mental status Abdominal pain
Major Surgical or Invasive Procedure	paracentesis reDACTED therapeutic paracentesis none. ReDACTED apex paracentesis no. paracentesis NO. PARCENTESES REDACTES therapeutic para centesi. Non REDACTE Paracentesis No. Apex Paracentesis Affa e e paracent. REDAC. diagnostic. non. Substance. REDACTED diagnose. E. A Paracent Ensis RCADTED. TABLE RED. no REDACED. Paracentesis REDACTED diagnostic paracentesis.
History of Present Illness	In the past week, she notes that she has been having worsening abd distention and discomfort. she denies easy bruising, melena, BRBPR, hemetesis, hemoptysis, or hematuria. pt has brief period of confusion. the 'ReDAC. INNR. C/b. Ascites. in the previous week. He has been. H/o. Co. ascites, hiv on ART, h/o IVDU, COPD, bioplar, PTSD. we report a case of traumatic abd cirrhosis.
Past Medical History	- HCV Cirrhosis: genotype 3a - HIV: on HAART, REDACTED CD4 count 173, REDACTED HIV viral load undetectable - COPD: REDACTED PFT showed FVC 1.95 (65%), FEV1 0.88 (37%), FEFmax 2.00 (33%) - Bipolar Affective Disorder - PTSD - Hx of cocaine and heroin abuse - Hx of skin cancer per patient report
Social History	REDACTED REDACTED REDACTED REDACTED
Family History	she only has one brother that she is in touch with and lives in REDACTED. She is not aware of any known GI or liver disease in her family. she is not in contact with most of them. She has a total of five siblings, but is not talking to many of them. GI and liver disease are not known in the family if she is a GI / liver disease. I'm not aware a one-year-old and is not. her family is in a. she has one, and only has in touch. lives in. He is not, and isn't in contact and is in the. they.
	VS: 98.1 107/61 78 18 97RA General: in NAD HEENT: CTAB, anicteric sclera, OP clear Neck: supple, no LAD CV: RRR,S1S2, no m/r/g Lungs: CTAB, prolonged expiratory phase, no w/r/r Abdomen: distended, mild diffuse tenderness, + flank dullness, cannot percuss liver/spleen edge REDACTED distension GU: no foley Ext: wwp, no c/e/e. + clubbing Neuro: AAO3, converse normally, able to recall 3 times after 5 minutes, CN II-XII intact Discharge: PHYSICAL EXAMINATION: VS: 98 105/70 95 General: in NAD HEENT: anicteric sclera, OP clear Neck: supple, no LAD CV: RRR,S1S2, no m/r/g Lungs: CTAB, prolonged expiratory phase, no w/r/r Abdomen: distended but improved, TTP in RUQ, GU: no foley Ext: wwp, no c/e/e. + clubbing Neuro: AAO3, CN II-XII intact ADMISSION PHYSICAL EXAM: VS: T98.1 105/57 79 20 97RA 44.6ko GENERAL: Thin chronically ill appearing woman in no acute distress HEENT: Sclera anicteric. MMM. no oral lesions HEART: RRR, normal S1 S2, no

Figure 2, patient summary

Users do not need to worry about restrictions on entering a subject id. Because when the user enters an invalid value, the screen will show "No information found for this patient, Please check the subject ID and try again." . This avoids unnecessary trouble.

The screenshot shows the same web application as Figure 2, but with an invalid subject ID. The sidebar on the left is identical, with the input field now containing "12345".

The main content area displays the title "Patient Data Summary Viewer" and a red error message:

No information found for this patient. Please check the Subject ID and try again.

Figure 3, invalid subject id input

In addition, the filter function provides the possibility for users to choose to read specific data. Based on the titles in the filter, the user can select the corresponding patient summary to view.

The 'select all' and 'clear all' buttons allow the user to select all and clear existing selections.

The screenshot shows a web application interface for viewing patient data. On the left, there is a sidebar with a search bar labeled 'Enter Subject ID' containing the value '10000032'. Below the search bar is a 'Submit' button. Further down is a 'Filter by Key(s):' section with a list of checkboxes for various data categories: subject_id, Sex, Service, Allergies, Attending, Chief Complaint, Major Surgical or Invasive Procedure, History of Present Illness, Past Medical History, Social History, Family History, Physical Exam, Pertinent Results, Brief Hospital Course, Medications on Admission, Discharge Medications, Discharge Disposition, Discharge Diagnosis, Discharge Condition, Discharge Instructions, and Followup Instructions. Below the filter list are 'Select All' and 'Clear All' buttons. At the bottom of the sidebar are 'Back to Main Menu' and 'Save as CSV' buttons. The main area of the application displays a table with two columns: 'Key' and 'Value'. The table contains the following data:

Key	Value
subject_id	10000032
Sex	F
Service	MEDICINE
Allergies	Percocet / Vicodin
Attending	REDACTED REDACTED REDACTED REDACTED
Chief Complaint	Worsening ABD distention and pain abdominal fullness and discomfort altered mental status Abdominal pain
Major Surgical or Invasive Procedure	paracentesis reDACTED therapeutic paracentesis none , reDACTED apex paracentesis no , paracentesis NO , PARCENTESES REDACTES therapeutic para centesi , Non REDACTE Paracentesis No . Apex Paracentesis Affa e e paracent . REDAC . diagnostic . non . Substance . REDACTED diagnose . E . A Paracent Ensis RCADTED . TABLE RED . no REDACTED . Paracentesis REDACTED diagnostic paracentesis .
History of Present Illness	In the past week , she notes that she has been having worsening abd distention and discomfort . she denies easy bruising , melena , BRBPR , hemetesis , hemoptysis , or hematuria . pt has brief period of confusion . the reDAC . INR . C/b . Ascites . in the previous week . He has been . H/o . Co . ascites , hiv on ART , h/o IVDU , COPD , bioplar , PTSD . we report a case of traumatic abd cirrhosis .
Past Medical History	- HCV Cirrhosis: genotype 3a - HIV: on HAART, REDACTED CD4 count 173, REDACTED HIV viral load undetectable - COPD: REDACTED PFT showed FVC 1.95 (65%), FEV1 0.88 (37%), FEFmax 2.00 (33%) - Bipolar Affective Disorder - PTSD - Hx of cocaine and heroin abuse - Hx of skin cancer per patient report
Social History	REDACTED REDACTED REDACTED REDACTED
Family History	she only has one brother that she is in touch with and lives in REDACTED . She is not aware of any known GI or liver disease in her family . she is not in contact with most of them. She has a total of five siblings , but is not talking to many of them . GI and liver disease are not known in the family if she is a GI / liver disease . I'm not aware a one-year-old and is not . her family is in a . she has one, and only has in touch . lives in . He is not, and isn't in contact and is in the . they .

Figure 4, filter and 'select all' & 'clear all'

Once used, you can click the 'back to main menu' button to return to the main page and allow users to query the new patient data summary again. When the user wants to end the application, users just need to click the Close button in the top right corner of the page to exit.

b. Save as a csv file

When users want to select specific content and save it locally, they can click the "save as csv" button, which can save the selected patient summary content as a csv file locally.

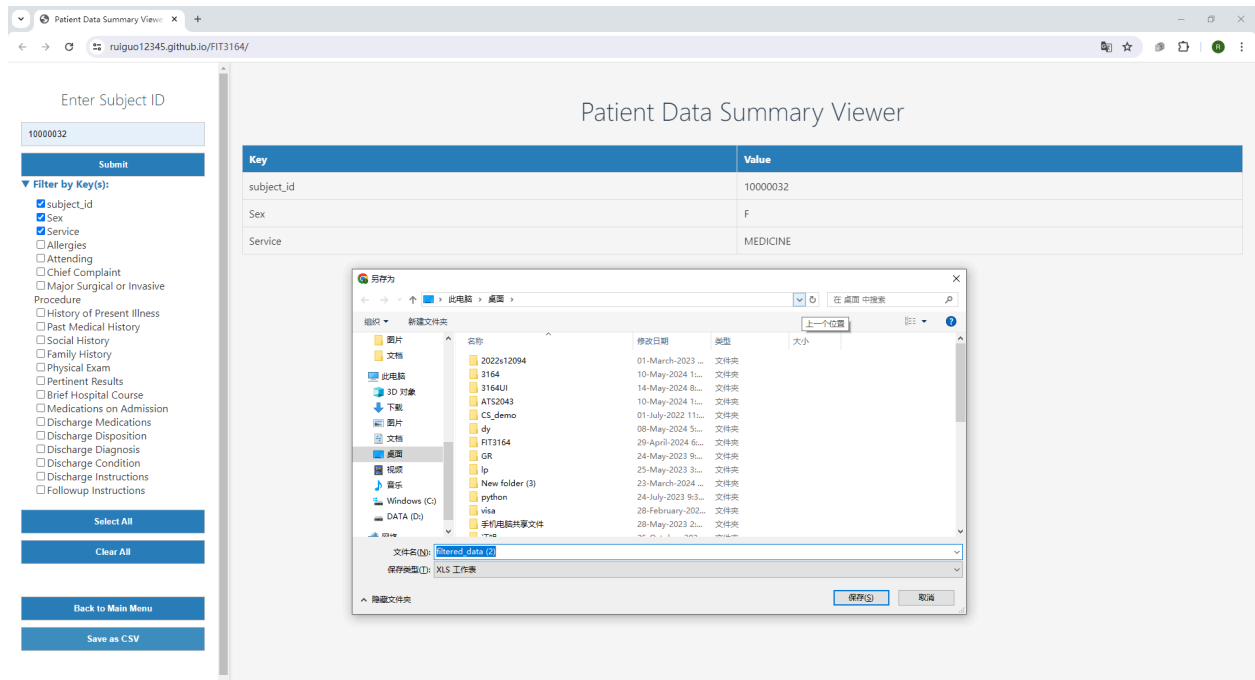


Figure 5, save selected data as a csv file locally

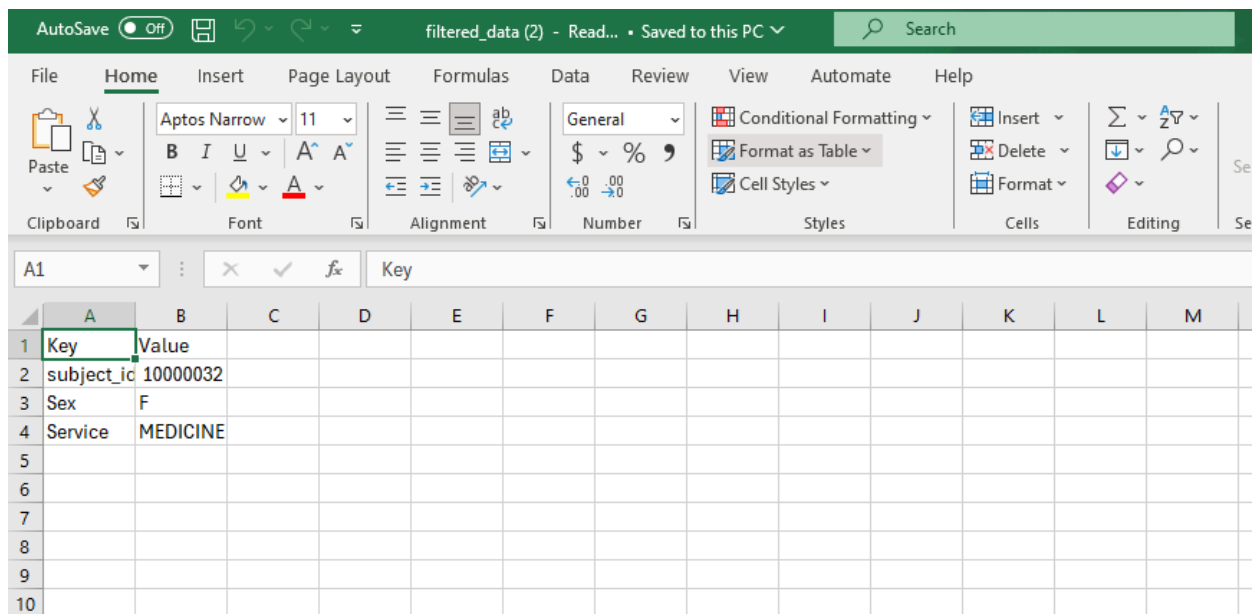


Figure 6, saved csv file

1.2. Limitations

a. Privacy of user

Since the users of this application may be doctors and patients, when a patient enters the id of another patient, he can still access the patient summary of this patient with this id. This issue involves personal privacy. It is recommended to use for doctors, or under the supervision or guidance of doctors to avoid leakage of patient information.

b. Update of data

Because the data in this application is the patient information after the LLM model summarized, and the connection between the database and the web page has not been completed. Therefore, when the data needs to be updated, the new data needs to be uploaded on github to access the latest patient data.

c. The name of saved files

When users save a csv file to the local PC, the default file name is filtered_data. After downloading multiple csv files, it may occur that users cannot distinguish between multiple csv files. Users should pay attention to timely renaming the files.

In addition, due to the default naming, users need to confirm the patient's subject id in time when downloading multiple patient records and add relevant information when naming locally to avoid confusion.

d. Accessibility

At present, the development of audio reading and other functions for disabled people has not been completed, and will be improved as soon as possible in the follow-up project work.

2. Technical Guide

The following is a user technical guide for the Patient summary application. This section describes the components of the application. The following sections describe the project details and the resources involved. This guide is intended to help teams who are not familiar with the project to facilitate subsequent program modifications, improvements, and maintenance.

2.1 Setup and Configuration

Before you start using the Patient Summary app, make sure you have the following software installed in registration and locally:

Requirement	Version	Link
Python	3.8 (at least)	https://www.python.org/downloads/
Visual Studio Code	latest	https://code.visualstudio.com/download
MySQL	8.4.0	https://dev.mysql.com/downloads/mysql/
MySQL Workbench	8.0.36	https://dev.mysql.com/downloads/workbench/
Chrome	latest	https://www.google.com/chrome/
GitHub	latest	https://github.com/
Excel	latest	https://www.microsoft.com/en-au/microsoft-365/p/excel-home-and-student/cfq7tc0hlkr?activetab=pivot:overviewtab

2.2. User Interface

a. Front-end code and relevant data

The front-end code can be cloned from the git repository (<https://github.com/RuiGuo12345/FIT3164>) to your local repository. You don't need to worry about access rights, etc. This github is public status and contains the html code for the website (a file called index.html) and the data set required for the web page (a document called summary.csv).

b. Run and Test

If you do not clone the codes to the local device, you can still through this link (<https://ruiguo12345.github.io/FIT3164/>) to run and test. A detailed description of the functionality of this website is mentioned in Section 1.1, and project limitations are mentioned in Section 1.2.

If you have cloned the code locally, you may be able to use visual studio code to modify, improve, and test.

Since the website is currently based on data in csv files, please put csv and html in the same path. Otherwise, the website may not work properly. Detailed descriptions of the data after LLM model processing and the code for converting it into csv documents are present in section 1.3.

c. Limitation and Further Improvement

Since this git repository belongs to one of our team members, if you want to push to this git repository after making changes on this project code or data, you may need to ask permission from our team members or assign the code to your local or git repository.

2.3 LLM (Large Language Model)

The code for the LLM can be found on my git repository (https://github.com/gcha0022/DS_project_2)

The LLM used can be applied by using python after installing certain packages using the pip command on bash:

huggingface-hub,Transformers,torch,torchvision,torchaudio,sentencepiece,protobuf and accelerate.

Then we need to login to our hugging face account using bash and a private access token(<https://huggingface.co/docs/hub/en/security-tokens>).

The model has been taken from Hugging Face and is a pre-trained LLM suitable for summarizing medical data.

a. Model Details

Model Used: Falconsai/medical_summarization

Reason for Choosing This Model: The model is specifically trained on medical data, making it highly suitable for our project, which deals with summarizing medical data.

b. Preprocessing Steps

Before applying the LLM, several preprocessing steps were performed to ensure the data was suitable for summarization. The preprocessing involved the following steps and techniques:

1. Data Size Reduction:

Given the initial large dataset, we reduced its size to make it more manageable.

2. Data Cleaning:

The following libraries were used for data cleaning:

```
import numpy as np
```

```
import pandas as pd
```

```
import re
```

3. Data Formatting:

The text data was reformatted into multiple columns using a custom parsing function. Below is the script used for reformatting:

```
def parse_text(text):
    # Split the text into parts based on the identified headings
    import re
    pattern = re.compile(r'(Name|Unit No|Admission Date|Discharge Date|Date of Birth|Sex|Service|Allergies|Attending|Chief Complaint|Major Surgical or Invasive Procedure)')
    parts = pattern.split(text)[1:] # Split and remove the first split since it's before the first heading

    # Create a dictionary to hold the extracted data
    data_dict = {}
    for i in range(0, len(parts)-1, 2):
        key = parts[i].strip()
        value = parts[i+1].strip()
        data_dict[key] = value

    return pd.Series(data_dict)

# Apply the function to the 'text' column
parsed_columns = df2['text'].apply(parse_text)

# Concatenate the parsed columns with any other existing DataFrame columns
df_final2 = pd.concat([df2, parsed_columns], axis=1).drop('text', axis=1) # Optionally drop the original text column
```

This was learned and implemented using the help of resources such as

(<https://www.freecodecamp.org/news/how-to-parse-a-string-in-python/>)

),(<https://www.geeksforgeeks.org/split-a-string-into-columns-using-regex-in-pandas-dataframe/>

[ame/](#)) and (<https://chatgpt.com/>) which was used to ask the GPT to debug the code and

implement on our dataset.

4. Merging and Cleaning Data:

The text data was merged by Subject ID and further cleaned using the following steps:

```
def merge_rows_by_subject_id(file_path):
    # Load the CSV file into a DataFrame
    df = pd.read_csv(file_path)

    # Define a custom aggregation function that concatenates strings with two newlines between them
    def custom_concat(series):
        return '\n\n'.join(series.astype(str))

    # Define the columns to take the last entry only
    last_entry_columns = ['Service', 'Sex', 'Family History', 'Allergies', 'Past Medical History', 'Discharge Disposition']

    # Create a dictionary for aggregation
    # Use 'last' for specific columns and 'custom_concat' for others
    aggregation_functions = {col: (lambda x: x.iloc[-1]) if col in last_entry_columns else custom_concat for col in df.columns if col != 'subject_id'}

    # Group by 'subject_id' and aggregate using the defined functions
    grouped_df = df.groupby('subject_id').agg(aggregation_functions).reset_index()

    return grouped_df
```

```
# Assuming merged_df is your DataFrame
def clean_text(text):
    """Clean placeholder text and standardize spacing."""
    if pd.isnull(text):
        return text # Return NaN as is
    text = re.sub(r'__+', 'REDACTED', text) # Replace placeholders with 'REDACTED'
    text = re.sub(r'\s+', ' ', text).strip() # Reduce multiple spaces to one and strip leading/trailing spaces
    return text

# Apply the cleaning function to each column in the DataFrame
for column in merged_df.columns:
    if merged_df[column].dtype == object: # Ensure the column is of type object (textual data)
        merged_df[column] = merged_df[column].apply(clean_text)
```

This was learned and implemented using (<https://stackoverflow.com/questions/46636080/merge-rows-within-a-group-together>), (<https://stackoverflow.com/questions/875968/how-to-remove-symbols-from-a-string-with-python>) and (<https://chatgpt.com/>) which was used to ask the GPT to debug the function and implement on our dataset.

c. Model Application

The LLM was applied to the preprocessed data using the following steps:

```

from transformers import pipeline

# Load the summarization model from Hugging Face's Transformers library
summarizer = pipeline("summarization", model="Falconsai/medical_summarization")

def summarize_text(text):
    """ Summarize the text using the loaded model, handling long texts by summarizing in parts if needed. """
    if len(text.strip()) == 0:
        return ""
    try:
        # Assuming the model can handle the entire text directly if not too long
        summary = summarizer(text, max_length=1024, min_length=150, do_sample=False)
        return summary[0]['summary_text']
    except Exception as e:
        print(f"Error summarizing text: {e}")
        return text

# List of columns to summarize
columns_to_summarize = [
    'Major Surgical or Invasive Procedure',
    'History of Present Illness',
    'Family History', 'Pertinent Results', 'Brief Hospital Course',
    'Discharge Diagnosis', 'Discharge Condition', 'Discharge Instructions'
]

# Apply summarization to each specified column and update the DataFrame directly
for column in columns_to_summarize:
    print(f"Summarizing {column}...")
    if merged_df[column].dtype == object: # Ensure the column is of type object (textual data)
        merged_df[column] = merged_df[column].apply(lambda x: summarize_text(x) if pd.notna(x) else x)

```

This model was applied using (https://huggingface.co/Falconsai/medical_summarization) and (<https://chatgpt.com/>) which was used to implement the LLM on specific given rows on our dataset.

d. Output

The summarized data was saved into a CSV file using the following command:

```
merged_df.to_csv('summary_df1.csv', index=False)
```

e. Challenges and Solutions

Challenges Faced:

Data Size: The initial dataset was too large to handle efficiently.

Solution: Reduced the data size before processing.

Data Cleaning: Ensuring the data was clean and standardized was challenging.

Solution: Used multiple libraries and custom functions for thorough data cleaning.

LLM Integration: Integrating the LLM and ensuring it summarized the text correctly required several iterations:

Solution: Used the Hugging Face Transformers library (https://huggingface.co/Falconsai/medical_summarization) and sought help from online resources, ChatGPT(<https://chatgpt.com/>) and tutors to debug and optimize the code.

2.4 Database

Considering development cost constraints and project development time constraints, we decided to set the database locally. Up to now (May 2024). We have completed the local database setup. However, the connection between the database and the website still needs to be solved. At present we can supply the database SQL codes. Codes can be found from this link. (<https://github.com/RuiGuo12345/FIT3164-databasecode>)

To set up the database locally, we recommend that you use MySQL workbench or use CMD directly.

First of all, download the latest MySQL. When downloading, you need to set the username and password, remember it, you will need to use it in the subsequent cmd command.

Open the mysql command line tool:

In the Windows Command Prompt, run the command:

`mysql -u userName -p`

Enter your password when prompted.

Once you have completed the above steps, you can create a new table directly in the workbench using the sql code provided in git.

(<https://github.com/RuiGuo12345/FIT3164-databasecode>)

Resource links for detailed steps:

1. Set Up a MySQL Database on Windows

[https://www.microfocus.com/documentation/idol/IDOL_12_0/MediaServer/Guides/html/English/Content/Getting_Started/Configure/ TRN_Set_up_MySQL.htm](https://www.microfocus.com/documentation/idol/IDOL_12_0/MediaServer/Guides/html/English/Content/Getting_Started/Configure/TRN_Set_up_MySQL.htm)

In addition, the MySQL development team did not optimize enough for the latest version of MySQL workbench. This always results in a warning when the workbench connects to the database. Please click 'continue anyway'.

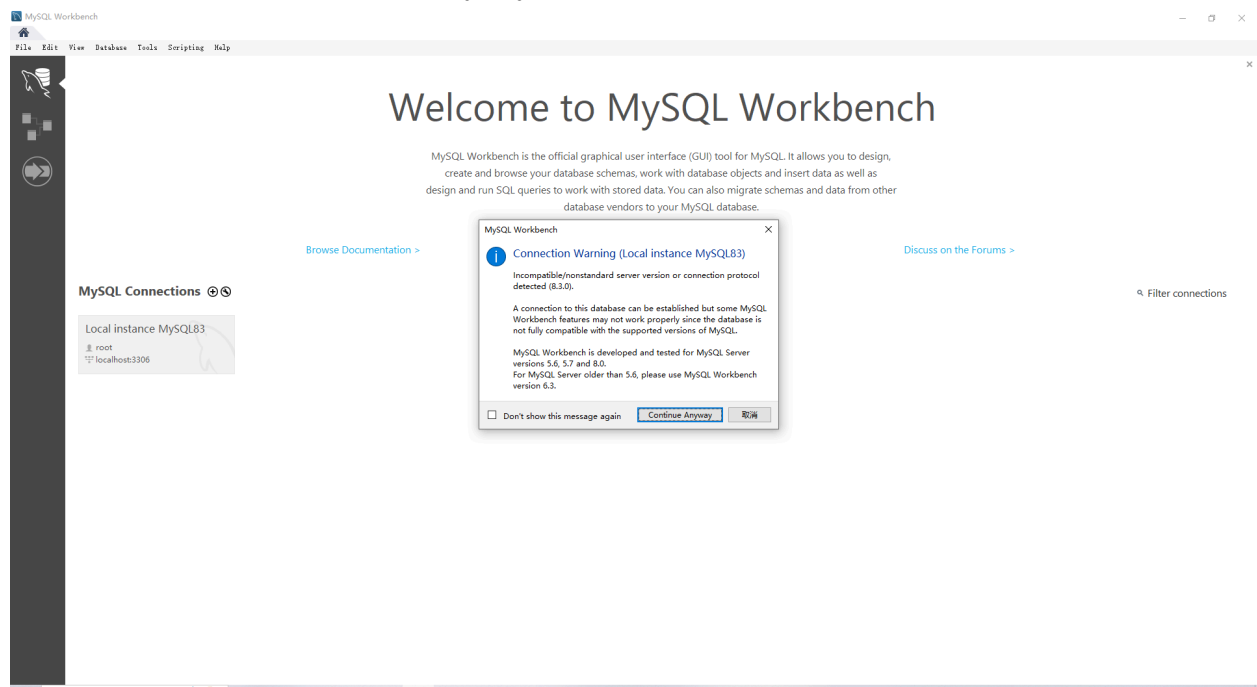


Figure 7, connection warning

3. Acknowledgement

We acknowledge the use of ChatGPT(<https://chat.openai.com/>) to organize language, check grammar, and make content understandable. The prompts used include checking grammar and improving our English writing. The output from these prompts was used to make content understandable and then avoid misunderstandings caused by our bad English Writing skills.